Notes on SDP Presentation Assignment

Audience: Individuals without a strong quant background

Primary Questions:

* Prevalence of dual enrollment course-taking in the state
* Characteristics of the students who take advantage of dual enrollment
* Association between dual enrollment and postsecondary success

Primary Metrics:

* Enrollment rates
* Remedial work in college

While dual enrollment does seem to influence

Questions I would like to address:

* Does the effect of dual enrollment differ by college type?
* What are the baseline factors that predicted dual enrollment?
* Did students who were dual enrolled do better on the ACT?
* Were students who were dual enrolled more likely to graduate on time?
* Were students who were dual enrolled less likely to take remedial coursework?
* Were students with more dual enrollment hours less likely to take remedial coursework?
* Were FRPL students less likely to take remedial coursework?
* Were students who were dual enrolled more likely to graduate?
* Do FRPL students benefit more from dual enrollment?
* Supplemental slide with distributions (kernel density plots) and categorical proportions,

Visualizations:

* For characteristics of students who take dual enrollment
  + Waffle charts
  + Proportional bar charts
* For looking at students who were dual enrolled moving into colleges
  + Alluvial plot

Future directions:

* Would be good to know whether students were first generation college students

Process:

* Checked out variables (e.g., Are there duplicate individuals? What are the range of values? Which variables have high levels of missingness?)
* Decided not to use ACT writing as it was highly missing (~95%)
* Releveled some of the categorical variables to have reference classes that were either the largest group or otherwise made theoretical sense as the reference
* Logistic regression to predict dual enrollment, because of the large number of observations decided to look at effect sizes and confidence intervals
* Visualized the variables that were the most relevant for comparison between dual and non dual-enrolled students
* Logistic regression to predict college enrollment, added ACT scores as predictors, left out sped because no students with IEPs went to college in this set
* Dominance analysis looks at change in R-square for adding/removing a variable once all the

Arguments:

* Predicting an outcome with a small baserate -- should be fine because there are many thousands of observations even though the overall baserate is low.
* Why not look at multi-level model -- it didn’t seem like school or district level would be very relevant across the entire state, seemed more relevant to look at a broader characteristic: school location
* Why not look at cohort -- seems unlikely that over a two year period we would expect major cohort effects
* Chick and the egg issue for predicting college outcomes -- were these students already advantaged; given that a number of control variables were included related to demographics, academic performance, and school environment, I believe this is a robust finding. however, it is possible that other confounding variables, such as parent education, more fine grained differences in parent SES,

Notes on stats:

Regression coefficient for categorical predictor is then the difference in likelihood of Y between the category for which X2 = 0 (the reference group) and the category for which X2 = 1 (the comparison group)]

Logistic regression for rare events shouldn’t be a problem so long as there is a very large sample, allowing for many instances of a particular outcome.

You can check the overall effect of the variable by performing a [likelihood ratio test](https://stats.stackexchange.com/questions/59085/how-to-test-for-simultaneous-equality-of-choosen-coefficients-in-logit-or-probit/59093#59093): fit two models, one with and one without the variable and type anova(model1, model2, test="LRT") in R (

Dominance analysis: This method is used to determine the relative importance of predictors in a regression analysis. Importance is defined as a qualitative comparison between pairs of predictors (Budescu, 1993). With this being said, one predictor is more important than another, if it contributes more to the prediction of the response variable in all possible subset models (i.e., all possible combinations of predictors) (Azen and Budescu, 2003). The McFadden index, *R2M*, is sometimes referred to as closer equivalent of the coefficient of determination, *R2*, in linear regressions.